机器学习中的数学第 6 课线性代数进阶

管枫

七月在线

June, 2016

主要内容

- 矩阵标准型
  - 把矩阵看做线性映射：相似变换
  - 把矩阵看做度量：相合变换
  - 正交相似变换
  - 奇异值分解
- 应用选讲
  - 主成分分析
  - SVD 在推荐系统中的应用
  - 正定矩阵与多变量凸函数
  - 极大似然估计渐进正态性

记号

- 本节课常用数学记号

V, W	向量空间
v, w	向量
$\mathbb{R}^n$,$\mathbb{R}^m$	实坐标空间
α,β	V 和 W 的基
$T : V \to W$	向量空间 V 到 W 的线性映射
$A_{\alpha,\beta}(T)$	线性映射 T 在 α 和 β 这两组基下的矩阵
$G(v_1, v_2)$	内积空间 V 上的内积
H_α	G 在基 α 下的矩阵形式

矩阵的标准型: 概述

矩阵的变换

- 标准型用来表示矩阵在变换下不变的性质
- 矩阵变换本质上是基的转换
- 相似变换：线性映射
- 相合变换：二次型（度量）

矩阵的标准型: 相似变换 (把矩阵看做线性映射)

如果 $T: V \to V$ 是一个线性变换, 那么对于 V 的两组基 α 与 $\tilde{\alpha} = \alpha \cdot P$，线性变换 T 的矩阵分别为

$$A_\alpha(T) \text{ and } A_{\tilde{\alpha}}(T) = P^{-1} \cdot A_\alpha(T) \cdot P$$

方阵的相似变换

- 如果两个方阵 A 和 $\tilde{A}$ 满足，$\tilde{A} = P^{-1}AP$. 那么这两个方阵就互为相似矩阵
- 相似矩阵的几何意义是同一个线性变换在不同的基下的表达形式
- 当研究对象是线性变换的时候，我们只关心矩阵在相似变换下不变的几何性质。

矩阵的标准型: 相似变换 (把矩阵看做线性映射)

矩阵的标准型: 相似变换 (把矩阵看做线性映射)

相似变换下不变的性质

- 行列式 (det)

$$\begin{aligned}\det(P^{-1}AP) &= \det(P^{-1})\det(A)\det(P)\\ &= \det(P^{-1})\det(P)\det(A)\\ &= \det(A)\end{aligned}$$

- 迹 (trace),$\mathrm{tr}(AB) = \mathrm{tr}(BA)$

$$\mathrm{tr}(P^{-1}AP) = \mathrm{tr}(APP^{-1}) = \mathrm{tr}(A\cdot I) = \mathrm{tr}(A)$$

- 秩 (rank)

矩阵的标准型: 相似不变量

相似变换下不变的性质

- 特征值：特征方程 $\det(A-\lambda I)=0$ 的根。
  如果 $\det(A-\lambda I)=0$，那么 $\det(P^{-1}(A-\lambda I)P)=0$，于是 $\det(P^{-1}AP-\lambda I)=0$

- 特征值是最重要的相似不变量，利用这个相似不变量可以方便的得出上面所有的不变量。

矩阵的标准型：相似不变量

Theorem (矩阵的相似标准型)

任何一个实系数方阵 A, 都存在一个可逆实系数方阵 P, 使得 $P^{-1}AP$ 是一个分块对角矩阵 *diag*$(J_1, ..., J_k)$. 且每一个对角块 (约当块)J_k 是如下四种情况之一：

i 1×1 伸缩变换矩阵块 $[\lambda]$

ii 2×2 伸缩旋转变换矩阵块 $R_{(\mu,\theta)} = \mu \cdot \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$

矩阵的标准型：相似标准型

iii $m \times m$ 循环伸缩矩阵块 $\begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix}$

iv $m \times m$ 循环伸缩旋转矩阵块

$$\begin{bmatrix} R_{(\mu,\theta)} & I_{2\times 2} & 0 & \cdots & 0 \\ 0 & R_{(\mu,\theta)} & I_{2\times 2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & R_{(\mu,\theta)} \end{bmatrix}$$

如果对角块中只有 [i] 和 [ii] 两种，那么这个矩阵称作可复对角化矩阵.

Proof.

I have discovered a truly marvellous proof of this, which this margin is too narrow to contain.– Fermat □

矩阵的标准型：相似标准型

Theorem (几乎所有方阵都可复对角化)

对于任何一个实系数方阵 A, 都存在一个可复对角化的矩阵序列 $\{A_i\}_{i=1}^{\infty}$, 使得

$$\lim_{i\to\infty} A_i = A$$

Proof.

证明思路:

1. 如果一个矩阵的特征多项式没有重跟，那么这个矩阵一定可以复对角化.
2. 几乎所有的矩阵特征多项式都没有重跟.

□

矩阵的标准型：相似不变量的几何意义

例题 (行列式 (det): 线性映射的体积膨胀系数)

如果 A 是线性变换 $T:V\to V$ 的矩阵，C 是 V 里边的立方体，那么：

$$\mathit{Volume}(T(C)) = \det(A)\cdot \mathit{Volume}(C)$$

Proof

因为行列式是相似不变量，不失一般性，可以假定 A 就是约当标准型.
证明分两步：先对可以复对角化的矩阵 A 进行证明，然后证明一般情况.
第一步：如果 A 是复对角化的矩阵，那么 $A = \mathrm{diag}(J_1, ..., J_k)$，其中 J_i 要么是一维的伸缩变换矩阵块，要么是二维的旋转变换矩阵块. 所以可以进一步把 A 写成

$$A = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_p, R_{(\mu_1,\theta_1)}, ..., R_{(\mu_q,\theta_q)})$$

于是 $\det(A) = \lambda_1 \cdot \lambda_2 \cdots \lambda_p \cdot \mu_1 \cdots \mu_q = \prod_{i=1}^{p} \lambda_i \cdot \prod_{j=1}^{q} \mu_j^2.$

Continue Proof

另一方面 $T(C)$ 是由 C 在前 p 个维度上进行拉伸，而在后面的维度上进行二维拉伸及旋转得到的. 所以

$$\text{Volume}(T(C)) = \prod_{i=1}^{p} \lambda_i \cdot \prod_{j=1}^{q} \mu_j^2 \cdot \text{Volume}(C)$$

所以

$$\text{Volume}(T(C)) = \det(A) \cdot \text{Volume}(C)$$

Continue Proof

第二步：对一般的 A，存在一个可复对角化的矩阵序列 $\{A_i\}_{i=1}^{\infty}$，及其所对应的线性变换序列 $\{T_i\}_{i=1}^{\infty}$, 使得：

$$\lim_{i\to\infty} A_i = A \text{ 而且 } \lim_{i\to\infty} T_i = T$$

根据第一步的结论, 对于任何 i 我们有,

$$\mathsf{Volume}(T_i(C)) = \det(A_i)\cdot \mathsf{Volume}(C).$$

因为体积与行列式都是连续函数，我们得到

$$\begin{aligned}\mathsf{Volume}(T(C)) &= \lim_{i\to\infty} \mathsf{Volume}(T_i(C)) \\ &= \lim_{i\to\infty} \det(A_i)\cdot \mathsf{Volume}(C) \\ &= \det(A)\cdot \mathsf{Volume}(C)\end{aligned}$$

于是证毕.

矩阵的标准型：相似不变量的几何意义

例题 (迹 (tr): exp(tr) 是线性映射 $exp(A)$ 的体积膨胀系数)

$$\exp(tr(A)) = \det(\exp(A))$$

Proof

首先对可复对角化的矩阵进行证明. 为了简化步骤我们不加证明的使用

Proposition

如果 A 是可复对角化矩阵，则存在一个复系数可逆方阵 Q, 使得 $\tilde{A} = Q^{-1}AQ$ 是复系数对角矩阵.

$$\tilde{A} = diag(\lambda_1, ..., \lambda_n)$$

Continue Proof

于是 $\exp(\tilde{A}) = \text{diag}(\exp(\lambda_1), ..., \exp(\lambda_n))$

$$\begin{aligned}\exp(\text{tr}(A)) &= \exp(\text{tr}(\tilde{A})) \\ &= \exp(\sum_{i=1}^{n} \lambda_i) \\ &= \prod_{i=1}^{n} \exp(\lambda_i) \\ &= \det(\tilde{A}) \\ &= \det(A)\end{aligned}$$

对于一般的 A, 可以模仿上一个证明中的第二步，留作思考题。

矩阵的标准型：相似不变量的几何意义

例题 (秩 (rank)：像空间的维数)

如果 A 是线性变换 $T: V \to V$ 在基 α 下的矩阵，那么

$$\mathit{rank}(A) = \dim T(V)$$

Proof

注意矩阵的秩并不是一个连续函数，所以对于这个问题我们不能仅仅考虑可复对角化的矩阵。我们必须直接考虑一般情况。
假定 $A = \mathrm{diag}(J_1, ..., J_K)$，其中每一个对角块都是约当标准型四中对角块里面的一种。于是我们得到空间 V 以及变换 T 的直和分解

$$V = V_1 \oplus \cdots \oplus V_k$$

$$T_i = T|_{V_i} : V_i \to V_i, \quad \text{for all } i \text{ from } 1 \text{ to } K$$

Continue Proof

所以

$$\dim T(V) = \sum_{i=1}^{K} \dim T_i(V_i)$$

而且

$$\mathsf{rank}(A) = \sum_{i=1}^{K} \mathsf{rank}(J_i)$$

所以要证明

$$\mathsf{rank}(A) = \dim T(V)$$

只需要对所有的约当块 J_i，证明

$$\mathsf{rank}(J_i) = \dim T_i(V_i)$$

对于每一个约当块的证明留作思考题。

矩阵的标准型

小结 (矩阵的标准型)

- 任何一个方阵总存在约当标准型（线性变换的相关问题总是存在一组好基）
- 如果问题具有相似不变性，可以假定矩阵是约当标准型从而简化问题
- 如果问题具有相似不变性以及连续性，可以假定矩阵是可复对角化的约当标准型，从而进一步简化问题

矩阵的标准型: 相合变换 (二次型)

假设 V 是一个实系数线性空间，那么线性空间上的度量指的是空间中向量的内积关系 $G(v_1, v_2)$. 如果 $\alpha\{\alpha_1, \cdots, \alpha_k\}$ 是空间 V 的一组基，那么这个内积一般可以用一个对称矩阵 $H_\alpha = [h_{ij}]_{n\times n}$ 来表示.

$$h_{ij} = G(\alpha_i, \alpha_j)$$

这时候对于任意两个向量 v_1, v_2, 如果 $v_1 = \alpha \cdot x_1, v_2 = \alpha \cdot x_2$，那么

$$G(v_1, v_2) = x_1^T H_\alpha x_2$$

矩阵的标准型: 相合变换 (二次型)

方阵的相合变换

- 如果两个对称方阵 A 和 $\tilde{A}$ 满足，$\tilde{A} = P^T AP$. 那么这两个方阵就互为相合矩阵
- 相似矩阵的几何意义是同一个内积结构在不同基下的表示形式

矩阵的标准型: 相合变换 (二次型)

相合不变量

- 矩阵的正定性（正定，负定）
- 矩阵的正负特征值的个数（Signature）
- 相合变换下矩阵保持对称性

我们涉及到的不定矩阵不多，所以关于相合不变量只需做大概了解

矩阵的标准型: 相合变换 (二次型)

方阵的正交相似变换

正交相似变换同时满足相似与相合变换的条件，也就是说它同时保持了矩阵的相似与相合不变量。

- 如果两个对称方阵 A 和 $\tilde{A}$ 满足，$\tilde{A} = P^T AP$. 而且 P 是正交矩阵:$P^T = P^{-1}$. 那么这 A 与 $\tilde{A}$ 就互为正交相似.

矩阵的标准型: 相合变换 (二次型)

矩阵的标准型: 相合变换 (二次型)

方阵的正交相似标准型

任何一个对称矩阵 A 都可以正交相似于一个对角矩阵 D.

总存在一个正交矩阵 P 使得, $A = P^T DP$.

方阵的正交相似标准型的几何意义

矩阵的标准型: 相合变换 (二次型)

主成分分析 (PCA)

PCA 的主要目的是降维，也可以起到分类的作用

- 当数据维度很大的时候，如果相信大部分变量之间存在线性关系，那么我们就希望降低维数，用较少的变量来抓住大部分的信息.
- 一般来讲做 PCA 之前要做 normalization 使得变量中心为 0，而且方差为 1.

比较广泛应用于图像识别，文档处理，推荐系统

矩阵的标准型: 相合变换 (二次型)

主成分分析 (PCA)

- 首先计算变量之间的协方差矩阵 Σ(利用样本)
- 找到 Σ 的正交相似标准型

正交相似标准性的求解由计算机完成，我们主要关心他的几何意义

矩阵的标准型:PCA 例子

推荐系统

如果一个旅游网站里面有 10000000 个注册用户，以及 40000 个注册酒店. 网站有用户通过本网站点击酒店页面的记录信息.$A = [A_{ij}]_{10000000 \times 40000}$,$A_{ij}$ 表示第 i 个用户点击 j 酒店的次数.

- 如何评价酒店之间的相似度?
- 给定一个酒店，请找出与它最相似的其他几个酒店?
- 如果要给酒店分类，有什么办法?

矩阵的标准型:PCA 例子

矩阵的标准型: 相合变换 (二次型)

长方矩阵的奇异值分解 (SVD)

对于任何一个矩阵 $B_{m\times n}$, 存在正交矩阵 $P_{m\times m}$,$Q_{n\times n}$. 使得

$$B = PDQ$$

其中 $D_{m\times n}$ 是一个只有对角元素不为零的矩阵.

证明

考虑 B^TB 与 BB^T 这两个对称矩阵

- B^TB 与 BB^T 拥有相同的特征多项式，所以拥有几乎相同的正交相似标准型
- $P_1^TB^TBP_1 = D_P$ 是 B^TB 的标准型
- $Q_1^TBB^TQ_1 = D_Q$ 是 BB^T 的标准型

那么考虑 $\tilde{B} = Q_1^TBP_1$, 我们知道

$$\tilde{B}^T\tilde{B} = P_1^TB^TQ_1Q_1^TBP_1 = P_1^TB^TBP_1 = D_P$$

另一方面

$$\tilde{B}\tilde{B}^T = Q_2^TBP_1P_1^TB^TQ_1 = Q_1^TBB^TQ_1 = D_Q$$

矩阵的标准型: 相合变换 (二次型)

继续证明

如果 $\tilde{B}$ 列数多于行数 $(m > n)$, 那么 $\tilde{B} = [B_1, O_{(n-m)\times m}]$. 而 B_1 的列向量彼此正交，长度平方为 D_Q 的对角元素. 于是存在正交矩阵 Q_2 使得 $Q_2^T B_1 = \sqrt{D_Q}$.

令 $D = Q_2^T \tilde{B} = Q_2^T Q_1^T B P_1$. 那么因为 $Q_2^T Q_1^T = (Q_1 Q_2)^T$ 仍然是正交矩阵，所以原命题得证.

矩阵的标准型:SVD 例子

推荐系统

如果一个旅游网站里面有 10000000 个注册用户，以及 40000 个注册酒店. 网站有用户通过本网站点击酒店页面的记录信息.$A = [A_{ij}]_{10000000\times 40000}$,$A_{ij}$ 表示第 i 个用户点击 j 酒店的次数.

- 如何评价用户的相似度?
- 给定一个用户的访问历史，请问他下一次最可能访问的酒店是哪一家?

矩阵的标准型:PCA 例子

多元函数的二阶导数

多元函数的二阶逼近

假设 x 是 n 维向量，$f(x)$ 是一个 n 元函数. 那么 $f(x)$ 在零附近的二次逼近可以写成

$$f(x) = f(0) + \nabla f(0)^T \cdot x + \frac{1}{2} x^T H(f)(0) x + o(|x|^2)$$

多元凸函数

如果 $H(f)$ 总是一个正定矩阵，那么 f 是一个凸函数. 多元凸函数仍然满足琴生不等式，而且我们在凸优化的课程中将会看到凸函数与凸集合的关系.

多参数的极大似然估计

考虑正态分布族 $N(\mu, \sigma)$. 两个参数分别是 (μ, σ). 假设 $X = (x_1, \cdots, x_n)$ 是 n 个样本，那么我们如何利用极大似然估计来进行参数估计呢?

$$L(\mu, \sigma) = (\frac{1}{\sigma})^n \prod_{i=1}^{n} exp(-(x_i - \mu)^2/(2\sigma^2))$$

$$l(\mu, \sigma) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - n \ln(\sigma)$$

多参数的极大似然估计

于是似然函数驻点方程为

$$0 = \frac{\partial}{\partial \mu} l(\mu, \sigma) = -\frac{1}{\sigma} \sum_{i=1}^{n} (\mu - x_i)$$

$$0 = \frac{\partial}{\partial \sigma} l(\mu, \sigma) = \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{1}{n\sigma}$$

求解得出估计值

$$\hat{\mu} = \overline{X}$$

$$\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{X})^2$$

极大似然估计的渐进正态性质 (optional)

Fisher information matrix

对于一个概率分布族 $f_\theta(x)$, 我们定义

$$\mathcal{I}(\theta)_{ij} = E(\frac{\partial}{\partial\theta_i}\ln(f_\theta(x))\frac{\partial}{\partial\theta_j}\ln(f_\theta(x)))$$

为 Fisher information matrix. 可以证明这个矩阵是半正定的.

极大似然估计当样本数量趋于无穷的时候的渐进分布为

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, I^{-1})$$

.

谢谢大家!